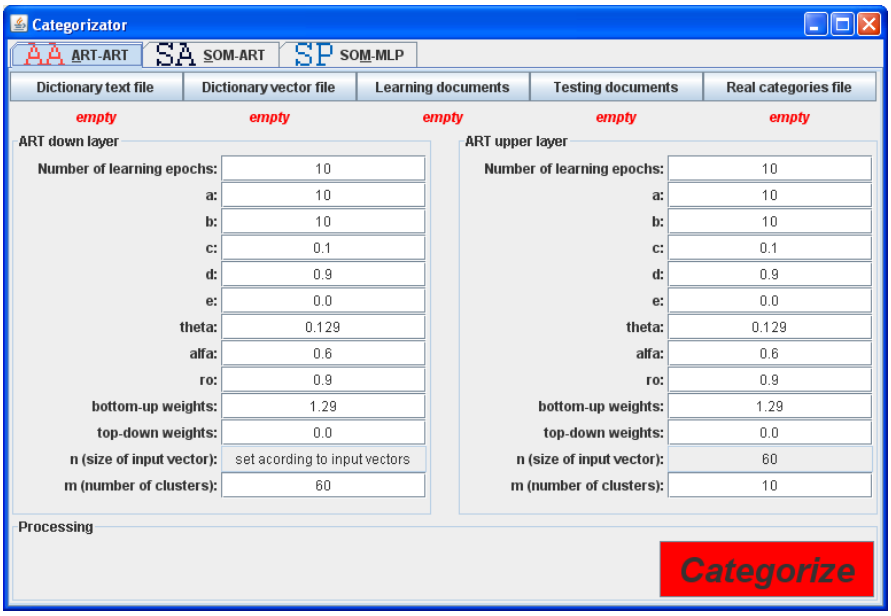
Uživatelská příručka

Aplikace Categorizator slouží ke kategorizování dokumentů třemi rozdílnými metodami. Každá z těchto metod podává jiné výsledky a to hlavně na základě zadaných parametrů. Před spuštěním tohoto programu je doporučeno nainstalovat JDK 1.6 pokud ho ještě na svém počítači nainstalované nemáte. Dále budeme potřebovat textové soubory, které lze získat pomocí bakalářské práce Lubomíra Krčmáře. Program je náročný jak na výpočetní výkon počítače, tak na paměť. Pro zpracování velkého počtu dokumentů (v řádech tisíců) je doporučeno aplikaci přidělit větší množství paměti přepínačem –Xmx.

Po spuštění aplikace se Vám uprostřed obrazovky objeví hlavní okno celého programu, které je zobrazeno na obr. 1.



Obr. 1 Hlavní okno aplikace – záložka ART-ART

Toto grafické uživatelské rozhraní je členěno na záložky. Hlavní okno obsahuje tři záložky a každá reprezentuje jednu metodu kategorizování textových dokumentů, tedy jednu kombinaci dolní a horní sítě. Následující kapitoly Vás provedou používáním každé kombinace.

ART-ART

Metoda ART-ART je zastoupena záložkou první. Než bude možné zpracování dokumentů spustit, je zapotřebí učinit několik kroků, které budou nyní popsány.

Načítání vstupních souborů

Nejvýše na této záložce můžeme vidět pět tlačítek, která nám umožní načíst všechny potřebné vstupní soubory. Tlačítko první je *Dictionary text file*. Kliknutím na něj vyvoláme zobrazení dalšího okna, v kterém je možné procházet všechna dostupná místa v našem

počítači a najít zde textový soubor, jenž obsahuje všechna slova ze všech dokumentů, které chceme kategorizovat. Druhé tlačítko nese název *Dictionary vector file*. Po jeho stisknutí se nám objeví stejné okno jako pro předchozí tlačítko, ale nyní budeme vybírat textový soubor, ve kterém se nachází slovní vektory každého slova celkového slovníku, tudíž toho souboru, který jsme načítali v předchozím kroku. Třetím tlačítkem je *Learning documents* a jak už název napovídá, půjde o načítání množiny učících dokumentů. Po stisknutí tlačítka se objeví již dobře známé okénko pro výběr souborů, ale tentokrát je umožněno vybírat více souborů najednou. Výběr souborů záleží čistě na uživateli, jen je nutné, aby všechny soubory měly za svým názvem písmeno *v*, které určuje, že obsahem souborů už nejsou dokumenty v textové podobě, ale jednotlivá slova dokumentů byla nahrazena kontextovými vektory. Následující tlačítko s názvem *Testing documents*, funguje obdobně. Jen se nyní do programu načítá testovací množina dokumentů. Dokumenty z této množiny budou po naučení sítí klasifikovány a výsledky klasifikace zobrazeny. Poslední tlačítko je popsáno *Real categories file* a slouží k načítání souboru, který obsahuje skutečné, tedy člověkem určené, kategorie dokumentů. Po každém načtení, které jsme pomocí pěti tlačítek v horní části záložky provedli, se objevil zelený výpis. Výpis je umístěn vždy pod tlačítkem, které načítání vyvolalo, a zobrazuje informaci o stavu. Při načítání samostatných souborů zobrazuje jejich název, při načítání množin ukazuje, kolik souborů množina obsahuje. Lze si tedy zkontrolovat, jestli nedošlo k omylu.

Nastavování parametrů

Další částí je nastavování parametrů sítí ART. Metoda ART-ART se skládá ze dvou ART sítí, horní a dolní. Parametry obou sítí lze nastavit. Už předem jsou všechna políčka vyplněna implicitními hodnotami, ale téměř každá z těchto hodnot lze změnit. Jedinou výjimkou je parametr ‘n’, jenž představuje délku vstupního vektoru. Délka vstupního vektoru dolní vrstvy ART je nastavena na základě vstupního souboru se slovními vektory a délka vstupního vektoru vrstvy horní je zase závislá na počtu neuronů ve vrstvě dolní, tudíž se na základě změny této hodnoty mění. Podrobný popis každého z parametrů a také jeho dovolené rozsahy hodnot lze nalézt v knize Fausett, L.: Fundamentals of Neural Networks, Prentice-Hall, New Jersey, 1994 na stránkách 218-287. Dle tohoto materiálu jsou také parametry pojmenovány a se stejnými jmény funguje celý program. Stručný výpis parametrů s jejich dovolenými rozsahy naleznete v tabulce 1. Význam a interval dovolených hodnot parametrů je pro obě vrstvy stejný.

Označení	Název/Význam	Interval povolených hodnot
number of learning epochs	počet učících epoch	<1 ; ∞)
a	a	<1 ; ∞)
b	b	<1 ; ∞)
c	c	(0 ; 1)
d	d	(0 ; 1)
e	e, brání dělení nulou	<0 ; ∞)
theta	parametr potlačování šumu	<0 ; ∞)
alfa	učící poměr (rychlost učení)	(0 ; 1)
ro	bdělostní parametr	<0,7 ; 1)
bottom-up weights	dopředné váhy	<0 ; ∞)
top-down weights	zpětné váhy	<1 ; ∞)
m	počet shluků	<1 ; ∞)

Tabulka 1 Parametry ART

Zpracování

V dolní části záložky ART-ART je umístěn blok s názvem *Processing*, ve kterém se nachází velké červené *Categorize* tlačítko. Jeho stiskem dáme programu pokyn ke zpracování zadaných vstupních souborů s uvedenými parametry. Během činnosti programu, protože může být v závislosti na velikosti zpracovávané množiny časově náročná, je zobrazeno malé okénko s nadpisem *Progress*. Úkolem tohoto okénka je zobrazovat stavové informace činnosti programu. Bude zobrazen název prováděné činnosti a procentuální vyjádření již vykonané části. Na závěr zpracování program informuje o dokončení své činnosti informačním oknem.

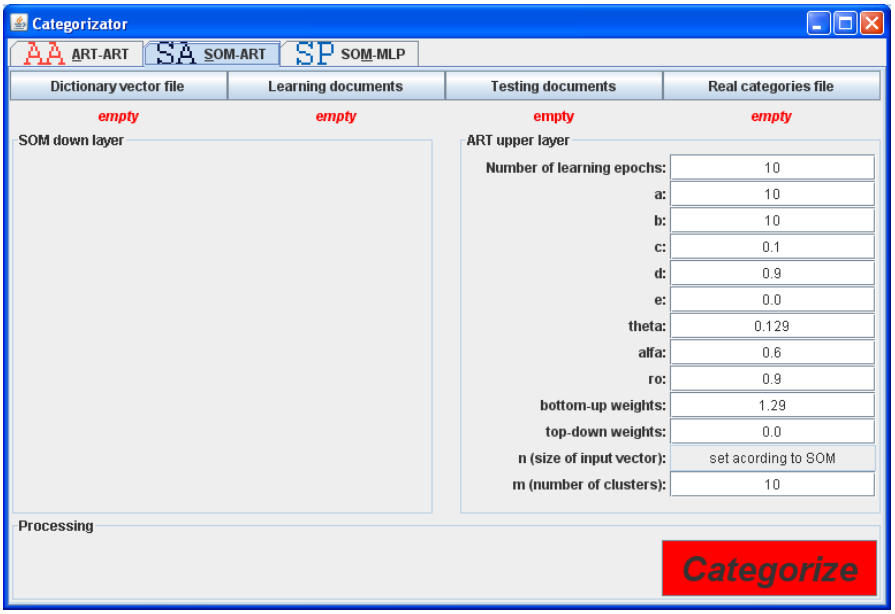
Výstup

Výsledky provedené kategorizace se nacházejí v adresáři na tom samém místě, kde byla aplikace spuštěna. Název adresáře je *Results yyyy-m-d h-m-s*, kde za klíčovým slovem *Results* je výpis data a času začátku kategorizace. Adresář obsahuje několik textových souborů. Soubory, jejichž název začíná slovem *Cluster*, které je následováno číslem, obsahují všechna slova, která byla zařazena do stejného shluku. Shluky neboli clustery jsou identifikovány čísly. Dalším textovým souborem je *Document vectors,* kde lze najít dokumentový vektor každého kategorizovaného dokumentu i s jeho skutečnou kategorií. Posledním souborem je soubor *Results yyyy-m-d h-m-s.txt*, kde datum a čas udává, kdy byla kategorizace dokončena. Tento soubor obsahuje všechny parametry, se kterými obě sítě ART pracovaly a výpis obsahu každého shluku kategorií. Shlukem se rozumí jeden cluster horní sítě ART. Výpis byl zvolen tak, aby i když do aplikace vložíme různé počty různých kategorií, dostaneme smysluplné a nezkreslené výsledky. Za číselným označením každého clusteru

následuje procento a název kategorie. Název kategorie je vzat ze souboru se skutečnými kategoriemi, který se načítal tlačítkem *Real categories file*. Procento, které tomuto názvu předchází, bylo získáno jako podíl počtu výskytů dokumentů z dané kategorie na tomto clusteru děleno celkovým počtem dané kategorie v celé databázi, samozřejmě vynásobeno stem, abychom dostali procenta. Tento výpočet pak dobře ukáže, kolik procent dokumentů z nějaké kategorie se řadí do jednotlivých clusterů. Pokud by se ve výsledcích objevila čísla přesahující 90 % v odlišných clusterech, pak by se dalo hovořit o úspěšné klasifikaci dokumentů. Pokud ale máme všude jen procenta malá nebo velká ale ve stejných clusterech, pak se zřejmě síti nepodařilo podchytit podobnost dokumentů ze stejné kategorie a odlišit je od kategorií jiných.

SOM-ART

Způsob zpracování SOM-ART lze vybrat kliknutím na druhou záložku v hlavním okně. Na obrazovce uvidíme okno z obr 2, kde lze nastavit vše potřebné pro kategorizaci dokumentů metodou SOM-ART.



Obr. 2 Záložka SOM-ART

Načítání vstupních souborů

V horní části okna jsou opět tlačítka pro výběr vstupních souborů. Nyní zde chybí první tlačítko pro výběr celkového slovníku, není zde třeba. Ostatní tlačítka fungují stejně jako u metody ART-ART a je také třeba všechny, zde čtyři, načítání provést.

Nastavování parametrů

Nastavení parametrů sítě SOM dovoleno není, ale parametry horní sítě ART lze nastavit a to opět stejně jako u předchozí metody. Opět se nevyplňuje parametr *n*, který

značí délku vstupních vektorů. Tuto činnost za nás vykoná program sám a tuto délku určuje na základě počtu neuronu vytvořených v síti SOM. Dovolené rozsahy ostatních parametrů se řídí tabulkou, která se nachází u návodu na ART-ART v části nastavování parametrů.

Zpracování

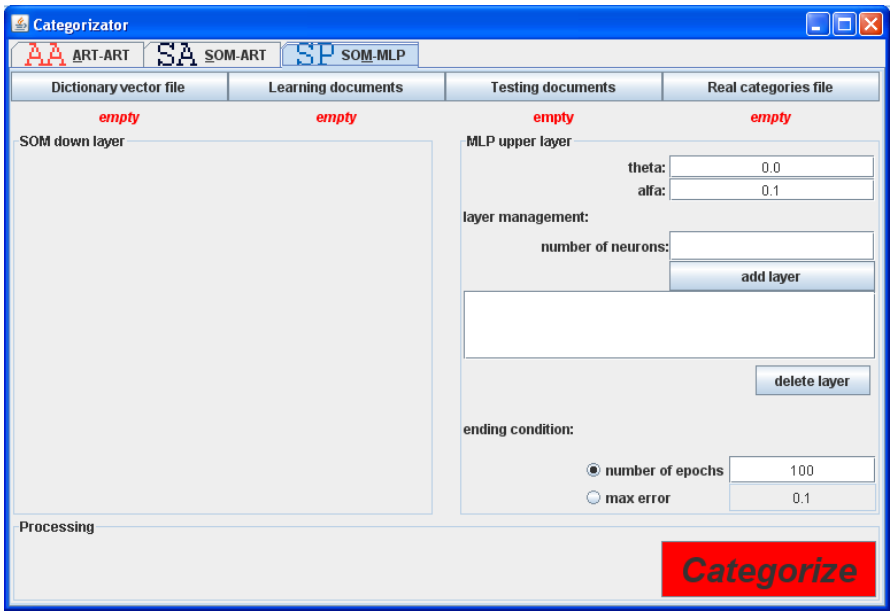
Zpracování se spouští tlačítkem *Categorize*. O průběhu zpracování stejně jako u kombinace ART-ART informuje okno *Progress*.

Výstup

Výsledky najdeme opět v adresáři pojmenovaném stejně jako u metody ART-ART, tedy *Results yyyy-m-d h-m-s*. To, že jsme kategorizovali jinou metodou, poznáme hned po otevření souboru s výsledky, který je také ve formátu *Results yyyy-m-d h-m-s.txt*, kde je na první řádce název použité metody napsaný. Na dalších řádkách se nachází výpis všech nastavených parametrů a na závěr výpis obsahů clusterů, který má stejný formát jako u metody ART-ART. Dokumentové vektory všech zpracovávaných dokumentů se nacházejí v souboru *Document vectors.txt*.

SOM-MLP

Třetí – poslední – záložka skýtá možnost zpracování dokumentů takovou kombinací sítí, kde SOM tvoří dolní a MLP horní síť. Záložka vypadá tak, jak ji zobrazuje obr 3.



Obr. 3 Záložka SOM-MLP

Načítání vstupních souborů

Tlačítka, která dovolí načtení příslušných vstupních souborů, jsou opět situována na horním okraji záložky. Jsou to ta samá tlačítka, kterými se načítá vstup u kombinace SOM-ART.

Nastavování parametrů

U sítě SOM se parametry opět nenastavují, nicméně síť MLP prostor pro nastavitelné parametry nabízí. Lze nastavit hodnotu parametru Θ a α. Zatímco Θ je parametr potlačování šumu, α udává učící poměr. Mimo nastavení těchto parametrů lze ještě modifikovat topologii sítě určením počtu skrytých vrstev a počtu neuronů v nich. Nová skrytá vrstva s žádaným počtem neuronů se přidá zadáním počtu neuronů do políčka s popiskem *number of neurons* a stiskem tlačítka *add layer*. Takto lze přidávat libovolné množství vrstev s libovolným počtem neuronů. Pro odebrání vrstvy ji stačí ve výpisu označit a stisknout tlačítko *delete layer*. Dále je možné si zvolit ukončovací podmínku. Lze ukončit zpracovávání po vykonání daného počtu učících cyklů, nebo po dosažení určené hodnoty chyby učení. Při výběru *number of epochs* se na hodnotu v poli *max error* nebere zřetel, ale při výběru *max error* slouží hodnota v poli *number of epochs* jako limitní počet epoch, po kterém se učení ukončí, i když chyby dosaženo nebylo. Nestane se tedy, že pokud nelze chyby dosáhnout, tak program nikdy neskončí.

Zpracování

Po zadání všech vstupních souborů lze spustit zpracovávání tlačítkem *Categorize*. O průběhu zpracování stejně jako u kombinace ART-ART a SOM-ART informuje okno *Progress*. Až dojde na učení sítě MLP, bude v okně *Progress* zobrazen graf s průběhem chyby učení.

Výstup

Výsledky obsahuje adresář *Results yyyy-m-d h-m-s*, s konkrétním datem a časem, kdy byla kategorizace započata. Tento adresář obsahuje soubor *Document vectors* stejně jako u všech metod zpracování. Soubor s výsledky s názvem začínajícím slovem *Results* má podobný obsah jako u sítí ART-ART a SOM-ART. Na jeho počátku jsou vypsány nastavené parametry a také informace o skrytých vrstvách. Síť MLP si vytváří tolik shluků, do kolika kategorií bude dokumenty klasifikovat a na rozdíl od ART-ART a SOM-ART jsou shluky rovnou pojmenovány dle kategorie dokumentů, které do daného shluku mají být řazeny. Výstupními soubory, které se u jiných kombinací sítí nenacházejí, jsou *errors.txt a outputs.txt*, přičemž první jmenovaný obsahuje vývoj chyby a druhý vývoj výstupů v průběhu učení.